DEMOGRAPHIC ATTRIBUTES BASED, COLD-START RECOMMENDATION OF MODULES IN ORGANIZATIONAL LEARNING

Peduru Hewage Suneth Dasantha Ekanayake

158213P

M.Sc. in Computer Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

June 2019

DEMOGRAPHIC ATTRIBUTES BASED, COLD-START RECOMMENDATION OF MODULES IN ORGANIZATIONAL LEARNING

Peduru Hewage Suneth Dasantha Ekanayake

158213P

This dissertation submitted in partial fulfillment of the requirements for the

degree Master of Science in Computer Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

June 2019

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

Signature: ..... Date: .....

Name : P.H. Suneth Dasantha Ekanayake

The above candidate has carried out research for the Masters Dissertation under my supervision.

Signature of the Supervisor: ..... Date: .....

Name : Dr. H.M.N. Dilum Bandara

ABSTRACT

Organizational learning is the process of creating, transferring, and retaining knowledge within an organization. It is of high importance due to the highly dynamic nature of the modern employee base. Moreover, new employees perform sub-optimally and get frustrated when such knowledge and expertise are not readily accessible to them. While many organizations use an organization-learning platform to bridge the knowledge gaps, both for new and existing employees, their effectiveness is being questioned due to lack of relevance, incoherent order of modules to be followed, and lack of fit with the learning style of an employee. While recommendation systems could overcome these challenges, it is difficult to provide a fitting set of recommendations for new employs who do not have any history with learning management system (aka., cold start problem).

We address the cold-start problem in recommender systems for organizational learning using the demographic information of employees. First, similar employees are grouped together based on their demographic attributes. Second, the modules that they follow are clustered according to their similarity. Then the orders of modules and the employee clusters are linked together in such a way that the number of module orders related to a user cluster is maximized. When a cold-start employee enters in to the system, his closest employee cluster is identified based on the demographic features and recommendations are generated considering the module sequences which have the least dissimilarities to the other module sequences in the linked module order cluster. We then tested the proposed technique using a synthetic dataset generated considering a medium scale organization. The dataset consists of age, gender, department, designation, and the order of learning modules followed by the employees. The proposed recommendation system has good accuracy, e.g., 71% of the module recommendations were more than 90% similar to the actual module orders.

Keywords: Collaborative Filtering, Order Clustering, Recommender System, Cold-Start Problem

ACKNOWLEDGEMENTS

My sincere thanks go to my project supervisor Dr. H.M.N. Dilum Bandara for guiding me throughout the project and giving a clear vision and a mission to accomplish during this period. It is essential to mention Mr. Sampath Tilakumara who is the Head of Technology at LSEG Technology, Sri Lanka for encouraging me on the research work.

The interest in machine learning was enriched by following the online learning materials in Coursera. Special mention has to be made regarding the Machine Learning module conducted by Prof. Andrew Ng of Stanford University in Coursera. The learning material and exercises followed in that course helped me a lot in this research work.

I take this moment to thank my parents for the support and the encouragement given to follow the MSc. in Computer Science amidst this busy schedule at working place. Last but not least, my sincere thanks go to the Head of Department of Computer Science and Engineering and the members of the academic staff for advising and guiding me during project evaluations and presentations.

TABLE OF CONTENTS

DECLARATION i

ABSTRACT ii

ACKNOWLEDGEMENTS iii

1 INTRODUCTION 0

Background 0

Problem Statement 2

Objectives 3

1.1

Contribution 4

1.2

Outline 4

1.3

1.4

2 LITERATURE REVIEW 5

1.5

Organizational Learning 5

Collaborative Filtering 6

2.1

2.2

Memory-Based Collaborative Filtering 7

2.2.1

Model-Based Collaborative Filtering 8

2.2.2

Hybrid Recommender Systems 8

2.2.3

Content-Based Filtering 8

2.3

2.4

Cold-Start Problem in Collaborative Filtering 9

2.4.1

User Cold-Start Problem 9

2.4.2

Item Cold-Start Problem 11

2.5

2.6

Clustering Algorithms 12

2.7

Clustering of Categorical Values 14

2.8

Clustering of Orders 15

3.1

Summary 17

3.2

3.23 METHODOLOGY 19

Overall Architecture of the Proposed Recommender System 19

Demographic Clustering of Users 20

Clustering of Numeric User Data 21

3.2.1.1 k-means Algorithm 21

Clustering of Categorical User Data 22
3.2.2.1 Dissimilarity Measure 22
3.2.2.2 Mode of a Set 22
3.2.2.3 k-modes Algorithm 23
3.2.2 3.2.2.4 Selection of initial modes 23
Clustering of Orders 23
Defining Order Clusters 24
Measuring the Similarity between Two Orders 24
3.3 Calculating the Order Mean 26
3.3.1 k-o’ means Algorithm 27
3.3.2 Order Prediction for Cold-Start Users 27
3.3.3 Explanation Using a Sample Dataset 27
3.3.4 Summary 35
3.4 4 PERFORMANCE EVALUATION 37
3.5 Synthetic Dataset Generation 37
3.6 Employee Demographic Information Dataset 37
4.1 Learning Module Preference Sequence 40
4.1.1 Evaluation Metrics 43
4.1.2 Experiment on Synthetic Dataset 43
4.2 RMSE Variation for the Synthetic Dataset 01 50
4.3 RMSE Variation for the Synthetic Dataset 02 52
4.3.1 Analysis of Results 54
4.3.2 Summary 56
4.3.3 5 CONCLUSIONS 57
4.4 Summary 57
Research Limitations 58
5.1 Future Work 59
5.2
5.3

LIST OF FIGURES

Figure 2.1: Categorization of recommender systems 7

Figure 2.2: Demographic based approach for new users 10

Figure 2.3: Conditional restricted Boltzmann machine with binary hide and binary visual units 12

Figure 3.1: Architecture of the proposed recommendation system 20

Figure 3.2: Order prediction algorithm. 28

Figure 3.3: User Clusters, k = 3 30

Figure 3.4: Clustering of learning module sequences, *k* = 3**Error! Bookmark not defined.**

Figure 4.1: Dissimilarities of the prediction and the actual, user cluster count=2, order cluster count=2 44

Figure 4.2: Dissimilarity distribution, user cluster count=2, order cluster count= 2 44

Figure 4.3: Dissimilarities of the prediction and the actual, user cluster count=2, order cluster count=5 45

Figure 4.4: Dissimilarity distribution, user cluster count=2, order cluster count=5 45

Figure 4.5: Dissimilarities of the prediction and the actual, user cluster count=3, order cluster count=2 46

Figure 4.6: Dissimilarity distribution, user cluster count=3, order cluster count=2 46

Figure 4.7: Dissimilarities of the prediction and the actual, user cluster count=3, order cluster count=5 47

Figure 4.8: Dissimilarity distribution, user cluster count=3, order cluster count=5 47

Figure 4.9: Dissimilarities of the prediction and the actual, user cluster count=4, order cluster count=2 48

Figure 4.10: Dissimilarity distribution, user cluster count=4, order cluster count=2 48

Figure 4.11: Dissimilarities of the prediction and the actual, user cluster count=4, order cluster count=5 49

Figure 4.12: Dissimilarity distribution, user cluster count=4, order cluster count=5 49

Figure 4.13: RMSE variation 50

Figure 4.14: RMSE variation 52

Figure 4.15: RMSE variation 56

LIST OF TABLES

Table 2.1: Rating distribution 16
Table 2.2: User Item rating matrix 16
Table 3.1: Demographic information of users 29
Table 3.2: Learning module sequences ordered according to the preference 31
Table 3.3: Dissimilarity values against each entry of cluster 0 32
Table 3.4: Dissimilarity sum values against each entry of cluster 0 33
Table 3.5: Common User Counts in User and Order Clusters 34
Table 3.6: Demographic information of the 41st employee 34
Table 3.7: Dissimilarity to user clusters 34
Table 3.8: Learning module sequence recommendations to the 41st employee 35
Table 4.1: Department wise employee distribution 38
Table 4.2: Designation wise distribution in Research and Development department 38
Table 4.3: Designation wise distribution in Sales and Marketing department 39
Table 4.4: Designation wise distribution in General and Administration department 39
Table 4.5: Designation wise distribution in Operations department 39
Table 4.6: Age categories 39
Table 4.7: Designation wise age distribution in departments 40
Table 4.8: Department wise module distribution 41
Table 4.9: Designation wise module completion percentage – mean value 42
Table 4.10: Learning module preference vector 43
Table 4.11: RMSE variation with user and order cluster sizes 50
Table 4.12: RMSE variation with user and order cluster counts 52
Table 4.13: Learning module sequence prediction and actual 55

LIST OF ABBREVIATIONS

Abbreviation	Description
CBF	Content-Based Filtering
CF	Collaborative Filtering
CRBM	Conditional Restricted Boltzmann Machine
MAE	Mean Absolute Error
RBM	Restricted Boltzmann Machine
RMS	Root Mean Square
RMSE	Root Mean Square Error